Business Analytics (110-1)

Assignment 2

Due: 9:00 am, Tue 02-Nov-2021

工管三 B08701119 張廷鋒

1.

GoodBelly is a company doing business mainly in the United States. We may see GoodBelly's sales outcome in different regions, in different stores, in different time periods, and with different marketing activities from the dataset "GoodBelly_data.csv". The data came from 126 Whole Foods stores over the 10 weeks between May 4 and July 13. There was a total of 1,386 observations.

The definition of the variables in the dataset are as follows.

1 Weekly Sales (Volume): The number of units sold per store per week.

2 Average Retail Price: The average retail price for GoodBelly products per store per week.

3 Sales Rep: Defined as 1 if the store had a regional sales rep (face-to-face contact) and 0 if the store had only the national sales rep (no face-to-face contact).

4 Endcap: Defined as 1 if a store participated in an endcap promotion.

5 Demo: Defined as 1 if the store had a demo on the corresponding week.

6 Demo1-3: Defined as 1 if the store had a demo 1-3 weeks ago.

7 Demo4-5: Defined as 1 if the store had a demo at least 4-5 weeks ago.

8 Natural Retailers: The number of other natural retailers within 5 miles of each store.

9 Fitness Centers: The number of fitness centers within 5 miles of each store.

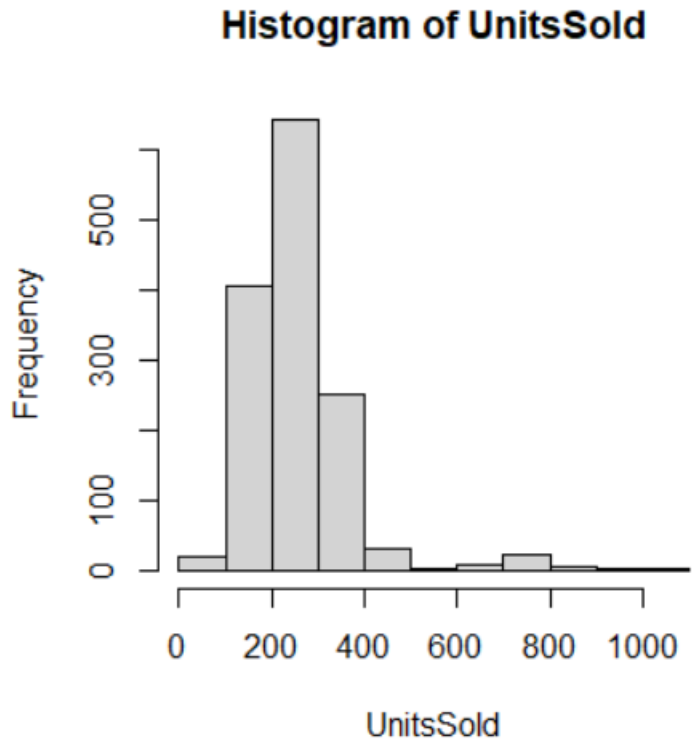
We would like to see whether those marketing efforts are effective and worthwhile. Investigate this data set and use the methods introduced in class to address the following questions.

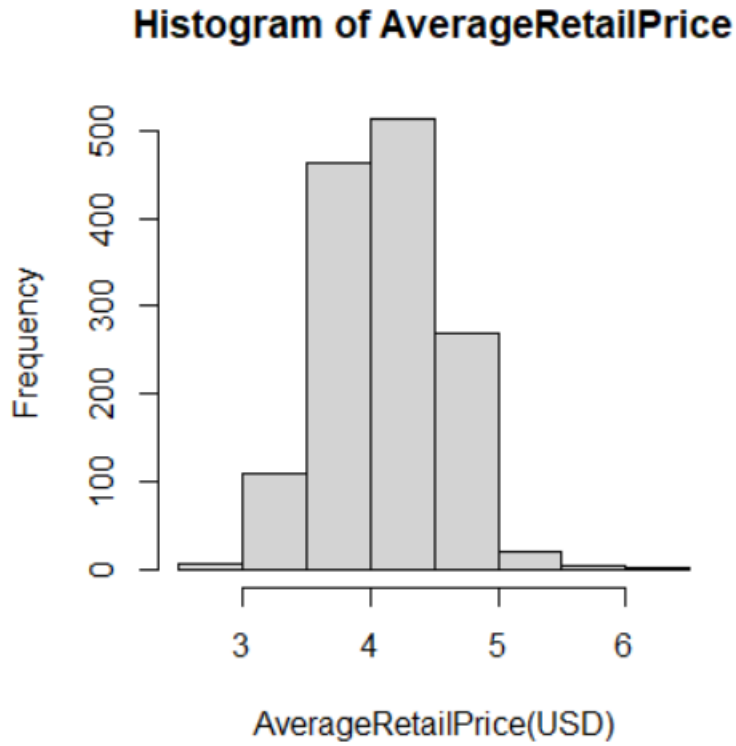
(a) Do descriptive statistics, with R, to provide an overview for your retailing business. You may do it from any perspective with any EDA method. You may include some basic summaries as well as some emphases on interesting findings.

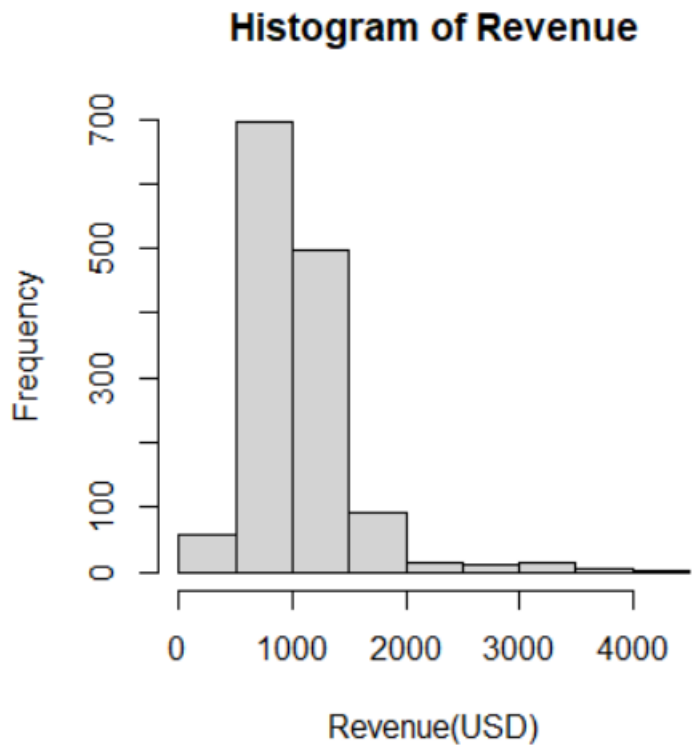
```
   UnitsSold       AverageRetailPrice    Revenue
 Min.   :  47.56   Min.   :2.889      Min.   : 204.2
 1st Qu.: 190.00   1st Qu.:3.776      1st Qu.: 762.5
 Median : 236.74   Median :4.097      Median : 967.3
 Mean   : 253.82   Mean   :4.107      Mean   :1041.5
 3rd Qu.: 295.80   3rd Qu.:4.425      3rd Qu.:1220.8
 Max.   :1041.20   Max.   :6.252      Max.   :4251.8
```

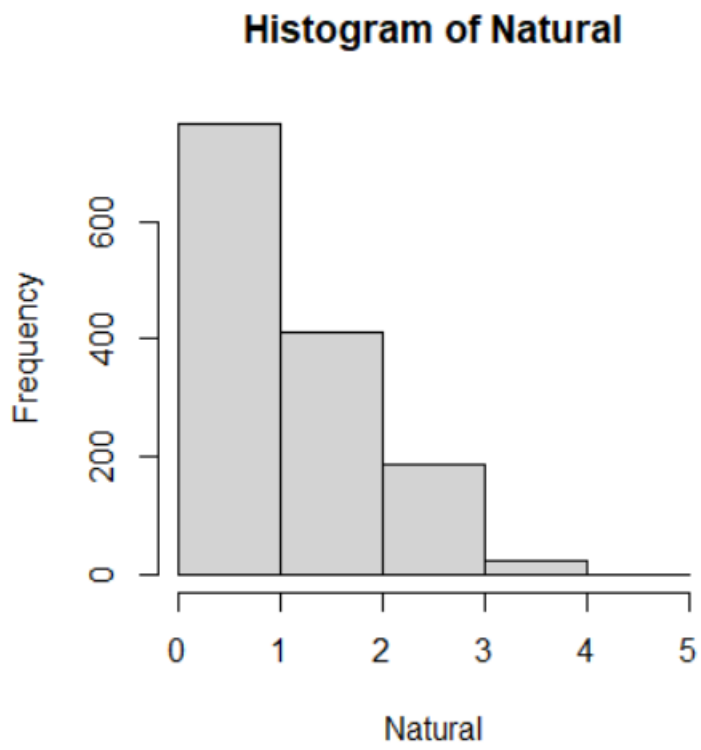
從簡易的summary得知，平均單日營收為1041.5，最高單日營收可到4251.8，最低則會低至204.2。

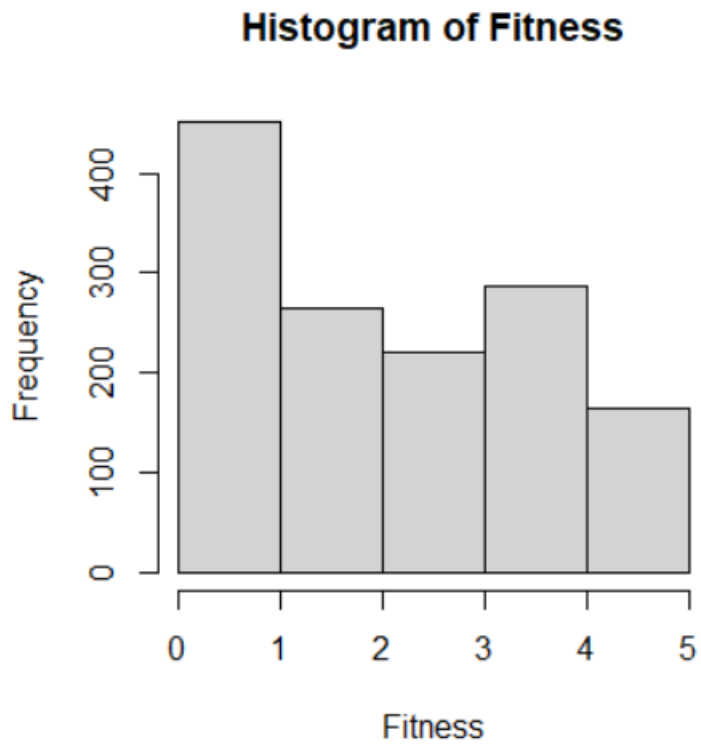
接著用histogram觀察各項數值變數分布情形。


Histogram of UnitsSold
Frequency
0
100
300
500
0
200
400
600
800
1000
UnitsSold

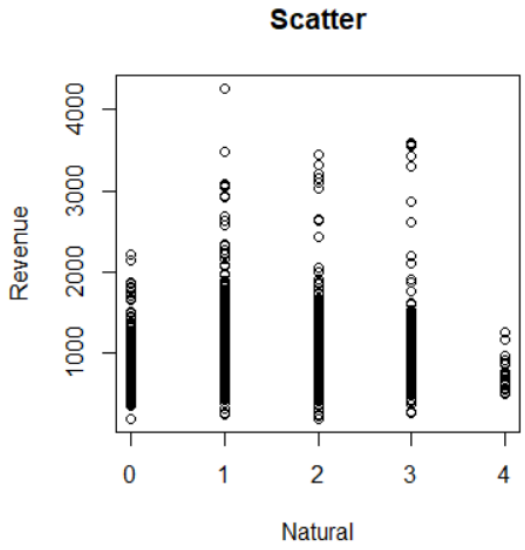

Histogram of AverageRetailPrice
Frequency
0
100
200
300
400
500
3
4
5
6
AverageRetailPrice(USD)

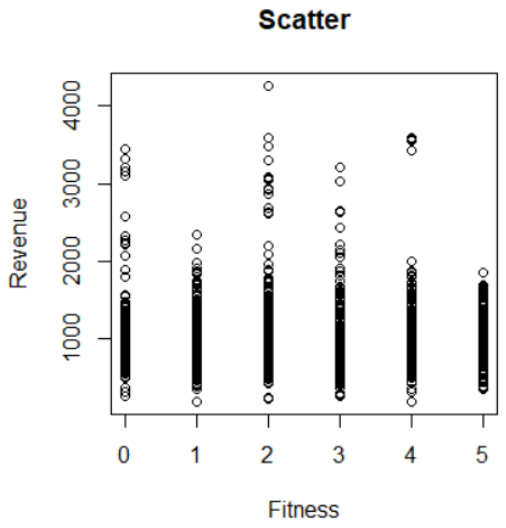

Histogram of Revenue
Frequency
0
100
300
500
700
0
1000
2000
3000
4000
Revenue(USD)


Histogram of Natural
Frequency
0
200
400
600
0
1
2
3
4
5
Natural

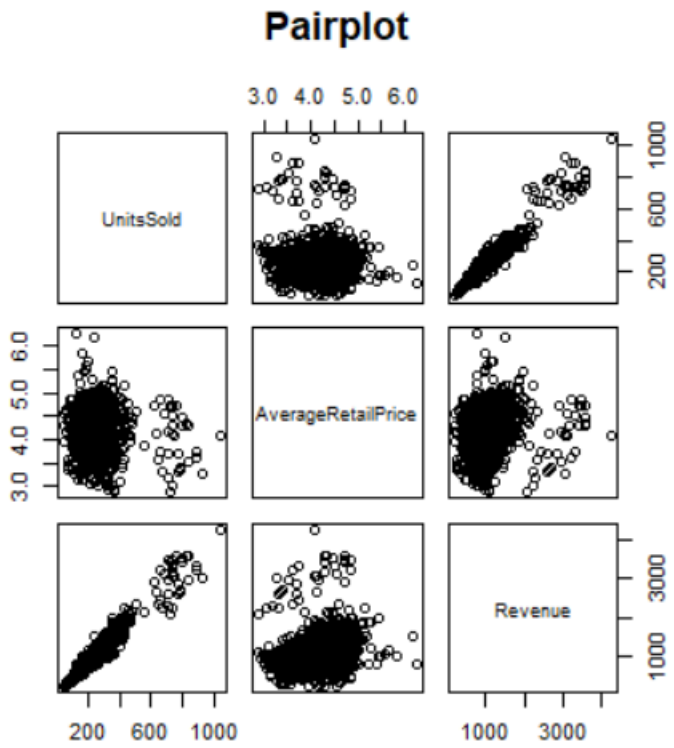


觀察這幾張histogram可發現，除了AverageRetailPrice以外其餘數值變數皆明顯左偏，若之後要建模型可能必須以log transformation轉換。
接著以pairplot觀察變數之間有無相關性。
由於Natural、Fitness值的種類較少且都為整數不具連續性，使用這兩個變數製作散布圖不具太大意義，故pairplot部分將略過這兩個變數。

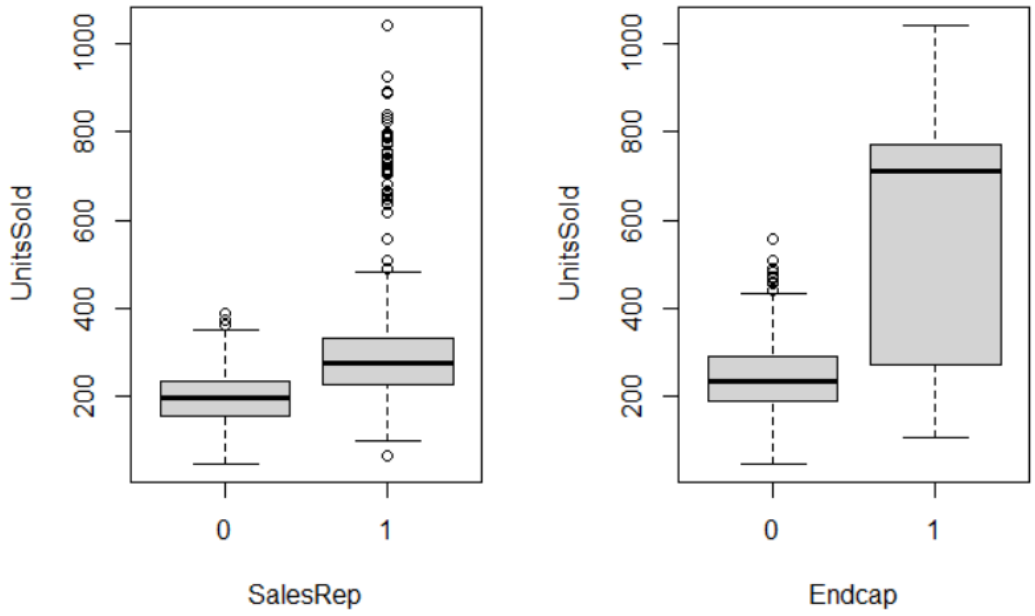




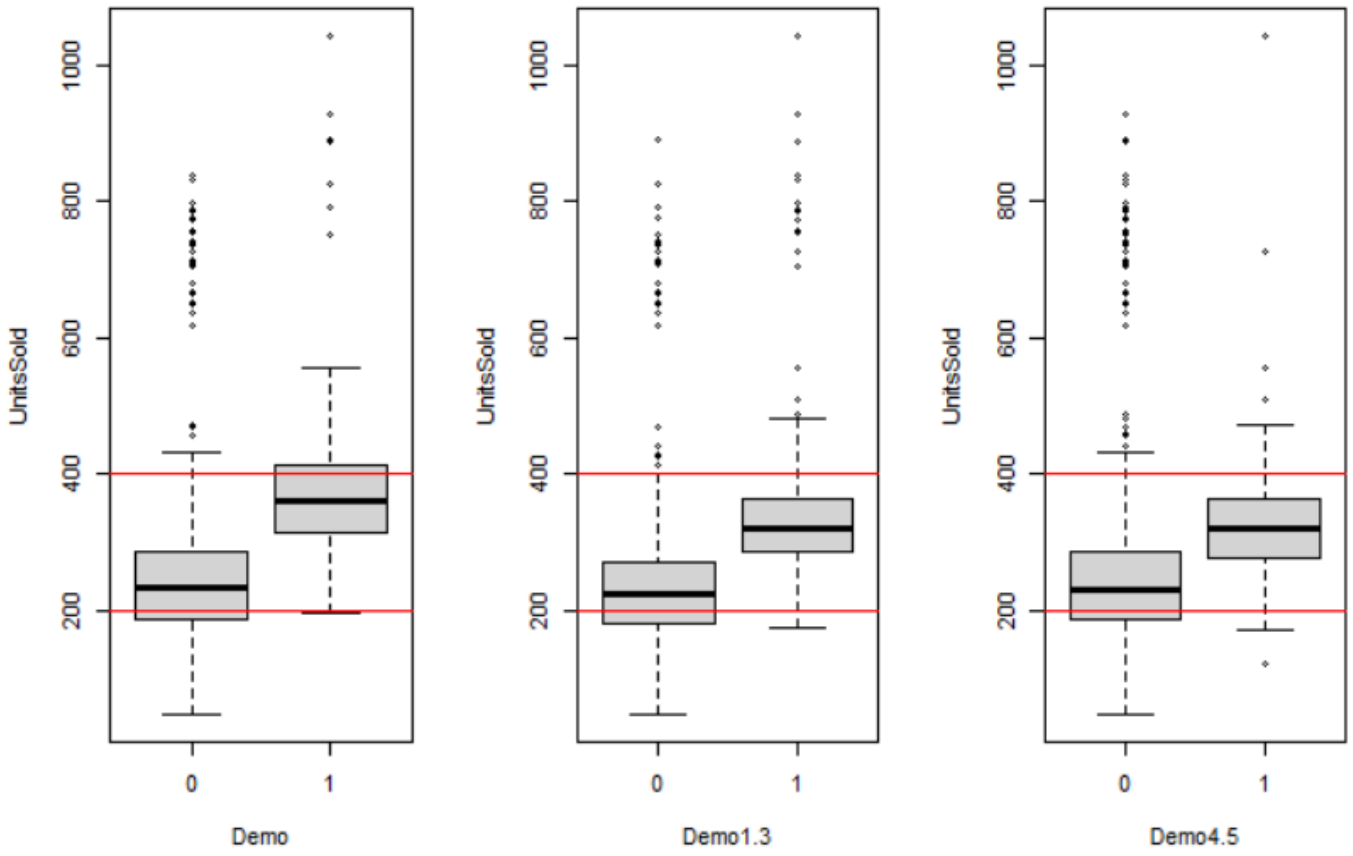
觀察Natural和Fitness對Revenue的散布圖並沒有發現太多資訊，故也不將其列入Pairplot中。



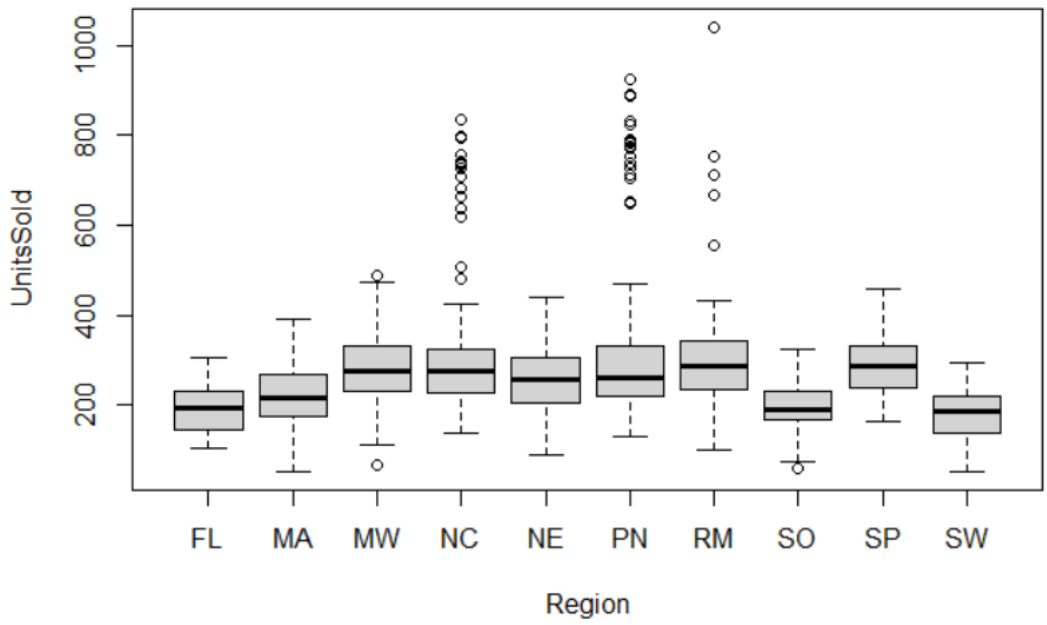
觀察pairplot後可發現，Revenue和UnitsSold正相關性較高，幾乎呈現一條直線的分布，和AverageRetailPrice則看不出明顯相關性。
總結以上觀察，我認為該公司的行銷活動可以注重於在合理的Price範圍內提升UnitsSold，因為UnitsSold幾乎和Revenue呈現完全正相關。



比較以上這張boxplot可以發現，Endcap對於UnitsSold的提升作用較SalesRep來的顯著許多。

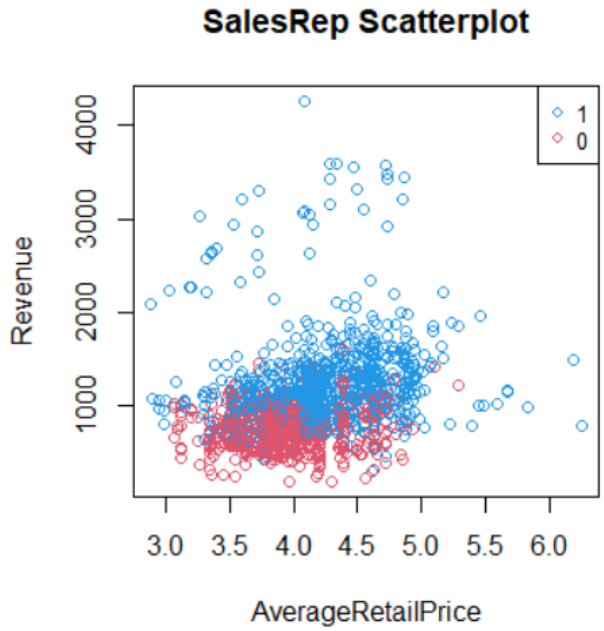


比較以上boxplot也可以發現，當周有Demo的情況下，銷售量的分佈有往上提升的效果，且此效果較1-3週、4-5週來的顯著。不過，銷售量的提升並不一定代表營收及利潤的提升，有可能只是廠商壓低價格吸引消費。

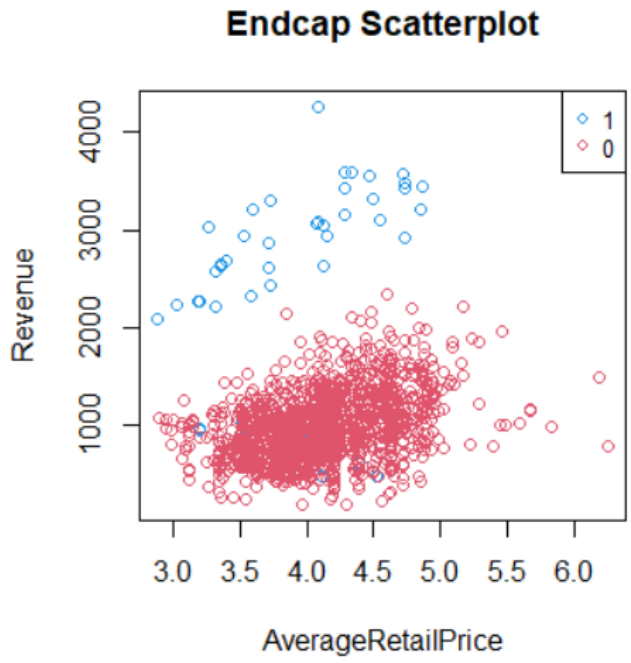


上圖為Region對UnitsSold的boxplot。可以得知公司的主力銷售地區為MW、NC、NE、PN、RM，於此五區UnitsSold分布皆較高。

(b) Based on your findings from (a), comment on the marketing activities about their effectiveness. Use some graphs and numbers to support your comments. You may comment on all of them, rank them, making suggestions about how to use them. Of course, your comments may be different from region to region, from time to time, or depending on any factor that you find useful.

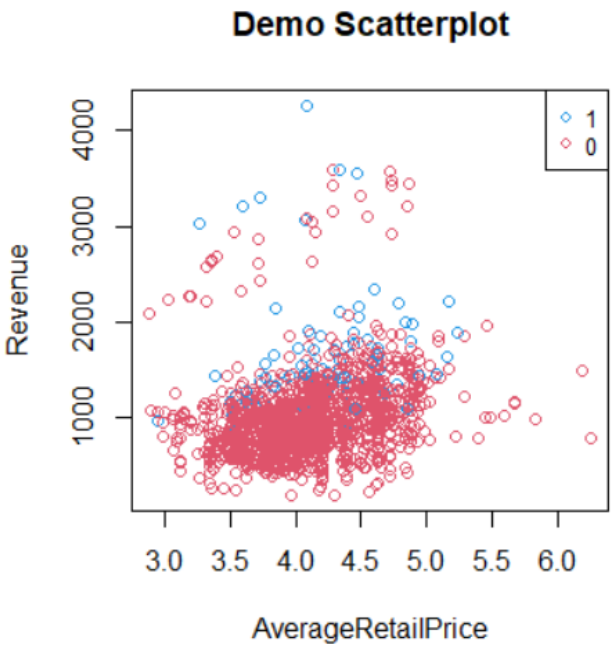


觀察以上這張圖可得知，有regional SalesRep提供Face to Face contact的門市營收通常比沒有的門市來得高，推測SalesRep具有提升營收的作用。此外，也可注意到有SalesRep的門市AverageRetailPrice有較多超過5.0的點，我推測部分擁有這樣資源的門市可能坐落在物價較高的城市地區。

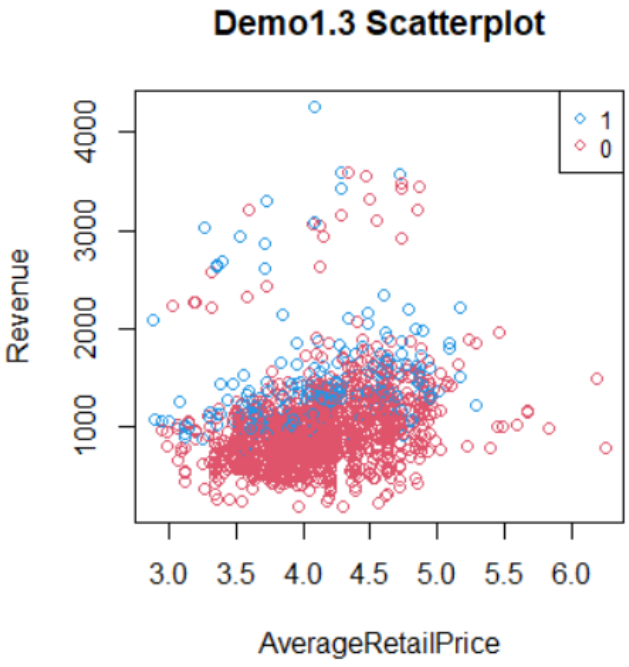


觀察以上這張圖可得知，有參加Endcap的門市AverageRetailPrice全部低於5.0，普遍較沒有參加的低，可見Endcap應是透過較大的價格優惠來吸引消費者。有參加endcap promotion的門市營收在這些時間點幾乎都比沒有參加的門市高，

推測Endcap具有顯著提升營收的作用。

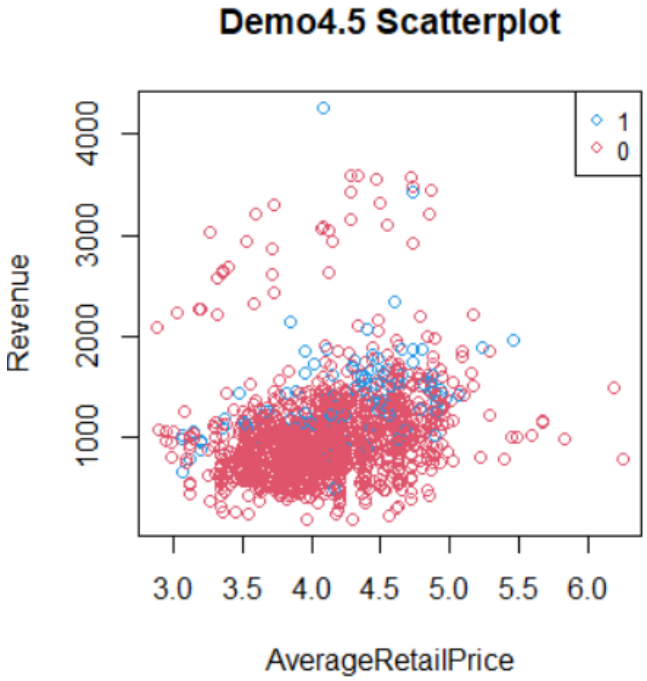


觀察以上散布圖可看到當周有發放Demo的門市Revenue僅略微高於當周沒有發放Demo的門市，但我認為這個差距可能不夠顯著，需透過統計方法證明這個行銷方式有效提升營收。



觀察以上散布圖可看到1-3周前有發放Demo的門市Revenue稍高於當周沒有發放Demo的門市，效果較當周發放Demo好。不過，依然無法用肉眼判定這個策

略夠不夠顯著，需透過統計方法證明這個行銷方式有效提升營收。



觀察以上散布圖可看到4-5周前有發放Demo的門市Revenue完全沒有高於當周沒有發放Demo的門市，可說是三個發放Demo的時間點中效果最差的。

若必須選擇一個時間點來發放Demo的話，我認為提前1-3週是一個較好的時間點。從資料的角度來看，這個時間點發放Demo帶來較高的營收；從消費者的角度來看，提前1-3週發放Demo給予消費者充分的思考與作購物規劃的時間，因此提前1-3週為最合適的時間點。

總結來說，我認為最有效的行銷方式是Endcap，其次是設立門市專屬的SalesRep，最後為發放Demo。

(c) Build a linear model to explain the relationship between sales and promotional efforts, and interpret the regression output.
下圖為ANCOVA公式

```
Call:
lm(formula = UnitsSold ~ (AverageRetailPrice + Natural + Fitness) *
    (SalesRep + Endcap + Demo + Demo1.3 + Demo4.5), data = data)
```

下圖最後一行為Global usefulness test結果，p-value = 2.2e-16 < 0.05，可以推翻H0: B0 = B1 = ⋯ = 0，證明此模型至少一參數不為0具有解釋力。

```
Residual standard error: 63 on 1362 degrees of freedom
Multiple R-squared:  0.6832,    Adjusted R-squared:  0.6778
F-statistic: 127.7 on 23 and 1362 DF,  p-value: < 2.2e-16
```

下圖為各參數p-value，以下將分為數值變數與類別變數分別闡述他們與銷售量

的關聯。

1. 數值變數

AverageRetailPrice	-31.6642	6.6907	-4.733	2.45e-06 ***
Natural	-7.4121	2.5648	-2.890	0.003914 **
Fitness	-1.2966	1.7185	-0.754	0.450689

(1) AverageRetailPrice：p-value為2.45e-06 < 0.05，可以推翻beta = 0的假設，證明此變數具有解釋力。斜率為-31.6642，代表在此模型中AverageRetailPrice每增加1，UnitsSold便減少31.6642。

(2) Natural：p-value為0.003914 < 0.05，可以推翻beta = 0的假設，證明此變數具有解釋力。斜率為-7.4121，代表在此模型中AverageRetailPrice每增加1，UnitsSold便減少7.4121。

(3) Fitness：p-value為0.450689 > 0.05，不能推翻beta = 0的假設，說明此變數不具有足夠解釋力。不過有可能與類別變數交互作用下產生更強的解釋力，因此先不將此變數從模型中剔除。

2. 類別變數—截距

觀察下圖結果可知，五個類別變數中，Endcap、Demo、Demo1.3的p-value < 0.05，可以推翻beta = 0的假設，證明這幾個類別變數在截距上與baseline model的差距是顯著的，也就是當他們為1時，截距會比baseline model(由於data中的類別變數level數都是2，故便是與類別變數為0時相比)分別上升377.43、184.52、88.62。

SalesRep	39.8606	35.3547	1.127	0.259751
Endcap	377.4258	70.3321	5.366	9.43e-08 ***
Demo	184.5168	66.8078	2.762	0.005824 **
Demo1.3	88.6187	40.0244	2.214	0.026986 *
Demo4.5	1.8331	55.6495	0.033	0.973727

3. 類別變數—斜率

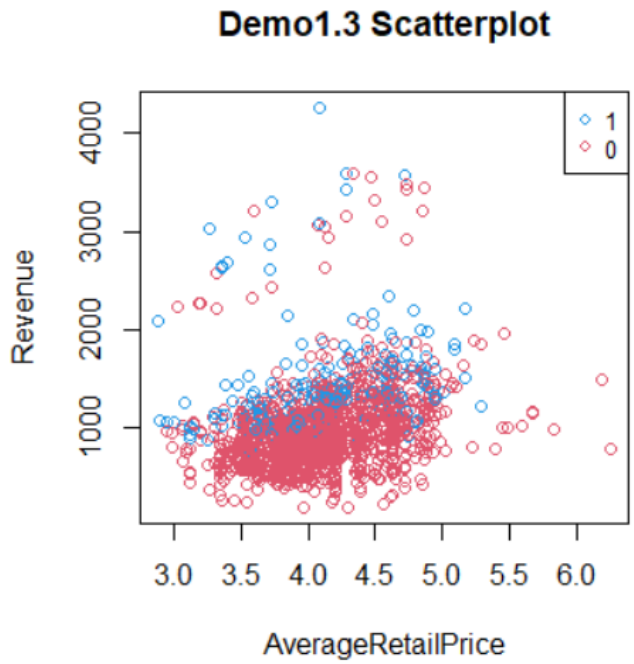
觀察下圖結果可知，只有Natural:SalesRep、Fitness:Endcap在斜率上與reference level產生顯著差距，不過這並不代表其他類別變數在斜率上完全沒有解釋力。

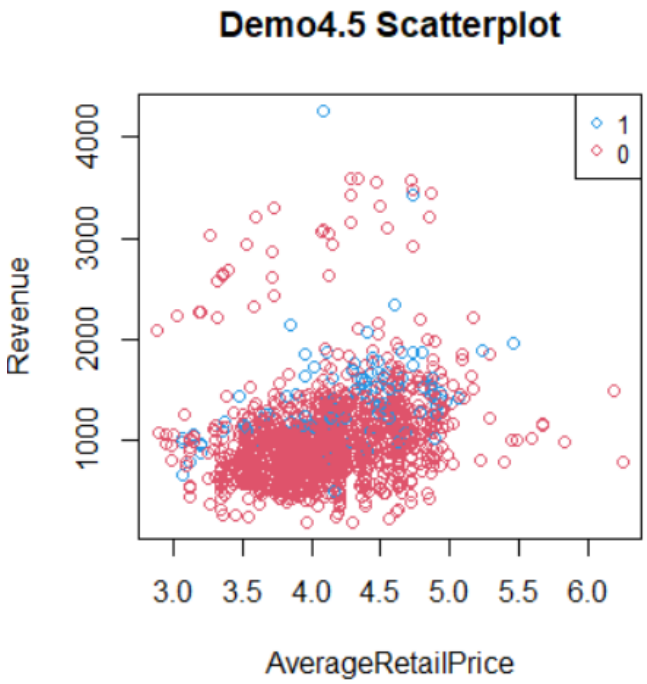
AverageRetailPrice:SalesRep	3.5890	8.4294	0.426	0.670339
AverageRetailPrice:Endcap	4.7893	17.1248	0.280	0.779772
AverageRetailPrice:Demo	-15.9185	15.5930	-1.021	0.307494
AverageRetailPrice:Demo1.3	-4.0364	9.5086	-0.424	0.671270
AverageRetailPrice:Demo4.5	14.8059	12.8442	1.153	0.249222
Natural:SalesRep	14.0963	3.7743	3.735	0.000196 ***
Natural:Endcap	-16.9965	11.8140	-1.439	0.150472
Natural:Demo	0.1582	7.9655	0.020	0.984153
Natural:Demo1.3	2.9141	5.3614	0.544	0.586851
Natural:Demo4.5	-1.4174	6.7906	-0.209	0.834695
Fitness:SalesRep	0.3148	2.2634	0.139	0.889392
Fitness:Endcap	-31.0835	7.5679	-4.107	4.24e-05 ***
Fitness:Demo	-1.2394	4.9332	-0.251	0.801675
Fitness:Demo1.3	0.3302	3.0903	0.107	0.914917
Fitness:Demo4.5	3.2780	4.2299	0.775	0.438503

由此model結果可以得知，數值變數部分與UnitsSold較具相關性的有

AverageRetailPrice和Natural，斜率皆為負呈現負相關。類別變數部分則是以SalesRep、Endcap、Demo、Demo1.3與UnitsSold較具相關性。

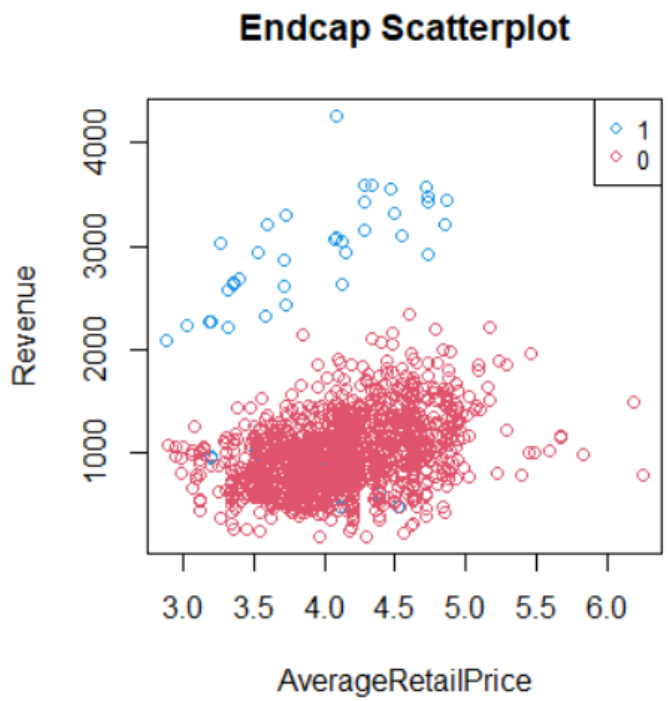
(d) Does the in-store demo program boost the sales? If so, for how long does the sales lift last?

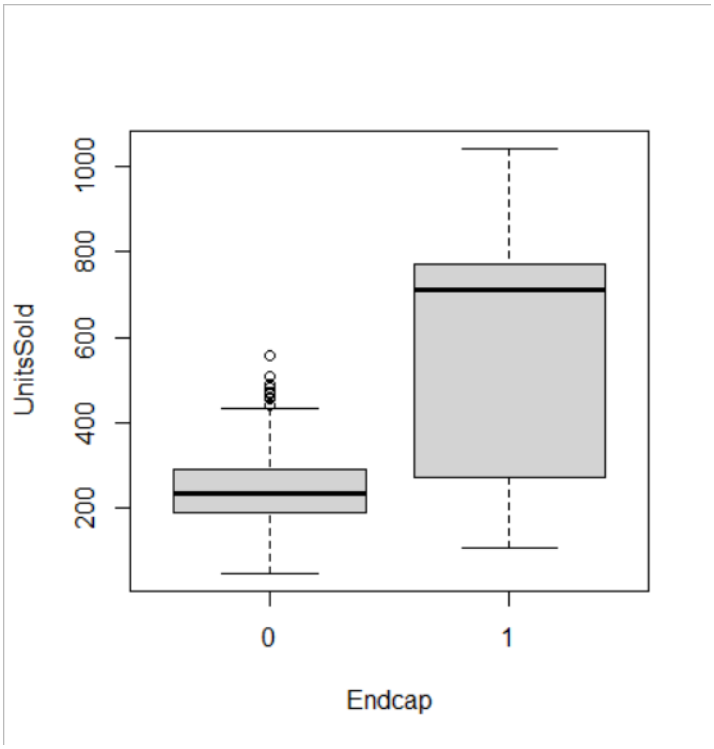




如同(b)小題中提到，我認為Demo對營收產生助益的時長約為1-3週，到4週以上的效果便和沒發放Demo的狀態相差無幾。

(e) Does the placement of the product within the store affect the sales?
從以下的scatter plot及boxplot便可以看出，Endcap為1的這些資料點在Revenue和UnitsSold的分布都明顯比Endcap為0的資料點高，有較明顯的區別。因此我認為Endcap確實對sales有影響。
此外，也可從model係數中觀察到，Endcap相關的係數p-value都常是顯著的。





(f) What other factors affect the sales of Goodbelly's products? Based on the regression output, what are your recommendations to Goodbelly's management?

1. Endcap promotion是所有行銷方法中最顯著提升營收的項目。
2. Demo具有提升營收的功效，但其效力最長只達三周，若欲使用Demo來提升營收須注意Demo的時效性。
3. UnitsSold和Revenue幾乎呈現完全正相關，並與AverageRetailPrice呈現負相關，可在適當價格範圍內舉辦特價促銷來提升營收。

(g) Are there any suggestions to improve and refine the model?

使用 stepwise regression 來移除對結果解釋力較低的變數。

下圖為 stepwise regression 後保留的變數，可以看到多數 p-value 都<0.05，部分>0.05 的項目則因為交叉項 p-value<0.05，對模型而言仍具有相當程度解釋力而被保留。

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                317.286     17.223  18.423  < 2e-16 ***
AverageRetailPrice         -31.238      4.088  -7.641 4.04e-14 ***
Natural                     -7.440      2.438  -3.052  0.00232 **
Fitness                     -0.956      1.103  -0.867  0.38603
SalesRep                    55.528      6.323   8.781  < 2e-16 ***
Endcap                     372.560     16.511  22.564  < 2e-16 ***
Demo                       114.125      7.323  15.584  < 2e-16 ***
Demo1.3                     76.441      4.864  15.717  < 2e-16 ***
Demo4.5                     -9.607     52.758  -0.182  0.85553
AverageRetailPrice:Demo4.5  18.774     12.363   1.519  0.12911
Natural:SalesRep            14.113      3.563   3.961 7.84e-05 ***
Fitness:Endcap             -34.003      6.975  -4.875 1.22e-06 ***
```

下圖呈現 stepwise regression 後的 R-squared 和 p-value。使用變數數量從 23 個下降到 11 個後，R-squared 並沒有下降太多，Adjusted R-squared 甚至上升，說明 stepwise 後的模型使用較少的變數達到同樣的效果，對公司來說穩定性上升也少了很多蒐集及維護資料的成本。

```
Residual standard error: 62.82 on 1374 degrees of freedom
Multiple R-squared:  0.6822,    Adjusted R-squared:  0.6797
F-statistic: 268.2 on 11 and 1374 DF,  p-value: < 2.2e-16
```